填补法和改进相似度相结合的协同过滤算法

邢长征，金　媛

(辽宁工程技术大学 电子与信息工程学院，辽宁 葫芦岛 125105)

摘　要：针对稀疏的用户评分数据，国内外学者对协同过滤算法做了很多改进，归纳为填充法、改进相似度方法、结合内容的推荐等，这些单一方法都不能真正解决数据稀疏的问题。针对这个问题，提出一种填充法和改进相似度相结合的协同过滤算法。该算法首先利用填充法随机填充部分数据，改进的填充法预测评分时融入了项目属性信息，然后利用填充后的数据和新相似度方法做推荐，产生推荐结果，迭代 m 次，按照迭代 m 次被推荐项目平均评分的高低进行最后的推荐。实验表明，在数据稀疏的情况下，该算法与单一的方法比有更好的推荐效果。

关键词：协同过滤算法；填补法；新相似度方法；结果融合

中图分类号：TP301.6　　doi: 10.3969/j.issn.1001-3695.2017.12.0813

Collaborative filtering algorithm combining filling and improving similarity

Xing Changzheng, Jin Yuan

(*School of Electronic Information Engineering Liaoning Technical University*, *Huludao Liaoning* 125105, *China*)

Abstract: Aiming at the sparse user rating data, domestic and foreign scholars have made many improvements on collaborative filtering algorithm, which were summarized as filling user rating data, improving similarity, fusing content to recommend and so on. These single methods can't solve the problem of data sparseness. In order to solve this problem, this paper proposed a collaborative filtering algorithm which combines the filling data and improving similarity. Firstly, it used the improved filling method which increases the item's attribute information to fill the user rating data, and then recommended using new similarity method, produced the recommended results, iterated m times. Finally it recommended items according to the average score of scores got in m iterations. The experiment shows that the proposed algorithm has a better recommendation effect than single methods in the case of sparse user rating data.

Key words: collaborative filtering algorithm; filling method; new similarity method; result fusion

0　引言

近年来，随着物联网，云计算，社交网络的迅速发展，网络空间所包含的信息量呈指数增长[1]。例如，亚马逊拥有数以百万种独特的产品，谷歌音乐库有数以千万计的歌曲，Del.icio.us 有超过 10 亿的网页收藏夹，淘宝的在线产品数量已超过 8 亿，新浪微博的用户和腾讯的微信用户超过 5 亿[2]。在这种情况下，推荐系统应运而生。推荐系统通过对用户的各种数据的收集和分析来学习用户兴趣和行为的模型，并向用户推荐所需的信息和服务。　推

荐系统能够有效解决信息过载问题，引起学术界和工业界的广泛关注。

Goldberg 等人于 1992 年开发了第一个推荐系统 Tapestry，并首次提出"协同过滤"思想[3]，这一思想的提出极大的推动了推荐系统的研究和发展。协同过滤算法根据用户的行为记录分析用户兴趣，找到与目标用户相似的邻居用户，综合这些邻居用户对某一信息的评价，形成系统对目标用户偏好的预测并进行相应的推荐[4]。协同过滤算法的优势在于：a）不需要考虑被推荐项目的内容，能够过滤机器难以分析的内容，如艺术品、电影、音乐等；b）有推荐新项目的能力，可以发现内容上不相似的项目，可以挖掘用户潜在的兴趣偏好；c）技术上容易实现。基于此，协同过滤技术是当前比较流行的推荐技术。然而，用户反馈信息矩阵是稀疏的，也就是说，大多数用户标记非常少的项目，导致传统的相似度计算方法不准确，难以获得较好的效果。

1　相关工作

1.1　协同过滤算法研究现状

协同过滤算法的基本假设是如果两个用户在一些项目上具有相似的历史标注或者行为习惯，那么他们在一些项目上也有

收稿日期：2017-12-21；修回日期：2018-02-02

作者简介：邢长征（1967-），男，辽宁省阜新人，教授，主要研究方向为数据挖掘和数据库（xcz6701@126.com）；金媛（1993-），女，硕士研究生，主要研究方向为数据挖掘.

相似的兴趣[5]。虽然协同过滤算法取得了巨大的成功，但仍然存在诸多问题，其中最为严峻的是数据稀疏问题。实际的网站中用户和项目的数目非常庞大，而多数用户通常只对小部分的项目评分，造成用户之间评分的重叠部分很小，难以计算两个用户之间的相似程度，并找到邻居用户，造成推荐结果不准确。针对这个问题，很多国内外学子提出了很多改进的方法，主要有空值填充法、改进相似度方法、推荐结果融合和结合内容推荐等方法。

1.2 缺失值填补法

缺失值填补法是根据已有的用户评分数据，以某种计算方法对用户未评分的数据进行估计并填充，可以显式的解决数据稀疏问题。最简单的填补法是将未评分的项目设一个固定的缺省值，或者设为其他用户对该项目的平均评分进行填充[6]。然而这种简单的填充法并不能满足用户的个性化需求，于是各种预测评分填补法被提出。邓爱林等人[7]提出基于项目评分预测的协同过滤算法，该算法采用基于项目的协同过滤方法填补确实数据。张玉芳等人[8]提出一种结合条件概率和传统协同过滤算法的非固定 K 近邻算法。该算法在基于分步填充评分矩阵的思想上，第一步只接受相似度和共同评分项目数量达到阈值的邻居用户作为目标用户邻居，然后计算并填充未评分项目；第二步使用第一阶段部分填充后的矩阵计算剩余未评分项目的评分。吕成戍等[9]提出了一个基于 KNN-SVM 的混合协同过滤推荐算法，该算法利用 K 最近邻法对训练集中的缺失数据进行填补,然后通过支持向量机交叉验证进行分类推荐。冷亚军等人[10]提出一种基于近邻评分填补的混合协同过滤推荐算法。该算法对原始评分矩阵进行全局降维，在低维的主成分空间中计算用户相似性，减少算法复杂度。采用奇异值分解法对近邻评分缺失值进行填补，降低近邻评分的稀疏性。Chujai 等人[11]同时使用用户信息和电影信息挖掘频繁项集，填补缺失数据。Insuwan 等人[12]提出 SVDUPMedianCF 算法，该算法利用改进的 K-means 算法进行聚类，得到聚类的中心来填补缺失值。

1.3 改进相似度方法

数据十分稀疏时，使用传统的相似度计算方法往往不能得到很好的推荐效果，于是研究人员提出很多新相似度计算方法。赵琴琴等人[13]提出一种改进的基于内存的协同过滤推荐算法 SPCF，该算法通过相似度传播，寻找到更多、更可靠的邻居。付芬等人[14]提出一种隐式评分和相似度传递的协同过滤推荐算法，该算法加入相似因子提高相似度的置信度，寻找最近邻用户，并引入相似度传递策略调整相似度因子产生推荐。仇国庆等人[15]设计了一种正态分布函数相似度量模型，此模型考虑了用户间的共同评分、共同评分项目数、以及用户的评分值，据此提出了融合正态分布函数相似度的协同过滤算法，该算法通过综合多种评分因素利用正态分布函数和修正的余弦相似度共同度量用户间的相似关系。Liu 等人[16]提出新的启发式相似性度量方法(new heuristic similarity model，NHSM)。NHSM 分别计算用户间的 PSS(proximity-significance-singularity)相似性、Jaccard 相似性和 URP(user-rating-preference)相似性，将三者乘积作为用户间最终的相似性。李容等人[17]考虑到用户共同评分项目占用比和平均评分因子作为两个修正因子来改进传统相似度的计算。

缺失值填补法可以直观、显著地改善数据稀疏问题，但它本身是对评分缺失值的一种预测，并不能真正代表用户偏好，而且预测的评分对推荐结果有较大的影响；改进相似度方法利用用户、项目、用户对项目评分等各种信息进行用户相似度的计算，但是它还是在已有的有限的评分数据集上进行计算，不能从根本上解决评分数据稀疏的问题。于是，本文提出一种填补法和改进相似度相结合的结果融合的协同过滤算法。该算法首先利用填补法随机填补部分数据，然后利用改进的相似度方法进行推荐，在原始数据上再填补部分数据，进行推荐，几轮之后，统计被评分项目的平均得分，按照得分多少依次推荐给用户。

2 本文算法

2.1 填补数据

本文首先对邓爱林等人[7]提出的基于项目评分预测的协同过滤方法的缺失数据填充部分作了改进。由于原有的填补方法只考虑到项目的评分，没有考虑到项目自身的属性，所以在原有算法的基础上加上项目属性的约束能使得缺失值填补更加准确。

设 $U=\{u_1,u_2,\cdots,u_n\}$ 是用户集合，$I=\{i_1,i_2,\cdots,i_m\}$ 是项目的集合，根据用户评分形成 *user-item* 评分矩阵 $R_{n\times m}$，对于用户未评分的项目默认设为 0。用户评分矩阵的列作为项目的特征向量，使用余弦相似度计算项目 i 和项目 j 之间的第一相似度，记为 $sim_{ij,1}$；查看各个项目与目标项目自身的属性信息，即对任意的项目 $j\in I$ 属性值与目标项目 i 的属性值相同，则第二相似度为 $sim_{ij,2}$ 等于 1，否则为 0。黄金比例已被广泛应用与建筑、美学、音乐、工业设计等领域，近年来，复杂系统的优化问题也借用了黄金比例，并取得了良好的效果[18]。本文借鉴“黄金分割”的思想，项目的相似度由两种相似度以黄金比例分割系数加权得到。由于用户对项目的评分更加直观体现用户的偏好，所以第一相似度对最终相似度的影响更大，即

$$sim_{ij}=0.618sim_{ij,1}+0.382sim_{ij,2} \tag{1}$$

按照相似度大小排序，形成目标项目的邻居项目集 M_p。根据公式

$$P_{ui}=\frac{\sum_{j\in M_p} sim_{ij}\times r_{uj}}{\sum_{j\in M_p} sim_{ij}} \tag{2}$$

预测用户 u 对目标项目 i 预测评分，并填充。

2.2 新相似度方法

在新相似性计算方法上，本文对李容等[17]提出的改进相似度的协同过滤算法中计算相似度的方法作出改进，下面列出改进的相似度计算方法的主要步骤：

a）考虑两个用户的共同评分项目数对相似度的影响

$$R(u,v)=\frac{n}{\min(N_u,N_v)} \tag{3}$$

其中，n 表示用户 u 和用户 v 的共同评分项目数，N_u,N_v 分别表示用户 u 和用户 v 的评分项目数。R 越大，则用户 u 用户 v 的整体相似度越高，正好符合两个用户共同评分的项目越多，两个用户的相似度越大的事实。

b）引入距离 $d(u,v)$ 来衡量用户 u 和用户 v 的平均评分差异。

$$d(u,v)=\frac{1}{n}\sum_{i\in I_{uv}}|r_{ui}-r_{vi}| \tag{4}$$

其中: r_{ui} 表示用户 u 对项目 i 的评分，I_{uv} 表示用户 u 和用户 v 的共同评分项目集。$d(u,v)$ 越大说明两个用户的平均评分差异越大，则两个用户相似性越低，则修正平均评分因子为：

$$p(u,v)=\frac{1}{1+d(u,v)} \tag{5}$$

p 越大则用户 u 和用户 v 的相似度越高。

c)得到改进的相似度计算方法：

$$NSim(u,v)=sim(u,v)\times R(u,v)\times p(u,v) \tag{6}$$

其中 $sim(u,v)$ 为传统的余弦相似度计算方法。

2.3 具体步骤

a）填补数据：为了避免用户对项目评分为 0 却被视为未评分的情况，在填补数据时，首先产生一个随机整数代表用户 *id*，计算用户评分项目的集合 I_u，项目集 $I-I_u$ 表示用户未评分的项目，再在 $I-I_u$ 里随机抽取一个整数作为项目 *id*，这样随机产生一个需要填补的数据。设定一个填补系数 $\alpha,\alpha\in[0,1]$，$\alpha=0$ 表示不对原始数据填补，$\alpha=1$ 表示对所有缺失数据进行填补，α 的大小决定填补的数据量。

b）填补数据后，使用新相似度方法，对未评分项目预测评分。

b）重复步骤 a）b）迭代 m 次，取每次被推荐项目评分的平均值，把项目按评分从高到低排序，依次推荐给用户。计算目标用户 u 对目标项目 i 预测评分 P_{ui}，公式为

$$P_{ui}=\frac{\sum_{v\in S}NSim(u,v)\times r_{vi}}{\sum_{v\in S}NSim(u,v)} \tag{7}$$

其中: r_{vi} 表示用户 v 对项目 i 的评分，S 是用户 u 的邻居集。

3 实验结果及分析

3.1 数据集

本实验使用 MovieLens 数据集[19]，该数据集由美国 Minnesota 大学 GroupLens 小组收集，包含了 943 位用户对 1682 部电影的 10 万条评分数据。所有的评分值分布在[0,5]区间，越高的评分值代表越强的用户兴趣，每位用户至少对 20 部电影评分。随机抽取 300 位用户的评分数据作为本实验的数据集，每组实验按照二八比例拆分数据集为测试数据和训练数据。

3.2 度量标准

为验证本文算法的性能，本文使用统计精度度量方法中最常用的评价指标平均绝对偏差(MAE)作为度量标准。平均绝对偏差 MAE 通过计算用户对项目的预测评分和实际评分之间的偏差度量算法的推荐准确性，MAE 越小，说明推荐结果越准确。设预测的用户评分集合表示为 $\{p_1,p_2,\cdots,p_n\}$，对应的实际用户评分集为 $\{r_1,r_2,\cdots,r_n\}$，则平均绝对偏差 MAE 定义为

$$MAE=\frac{\sum_{i=1}^{n}|p_i-r_i|}{n} \tag{8}$$

3.3 实验结果及分析

本节设计两组实验对本文算法进行分析研究。第一组实验分析填补系数 α 和迭代次数 m 对算法的影响，找到使算法推荐结果最优的填补系数 α 和迭代次数 m。算法最优的基础上，第二组实验使本文算法与其他推荐算法比较推荐效果。

首先确定迭代次数 m 的大小。设置填补系数为 0.2，每次迭代时，填充百分之二十的数据，推荐时为每个用户选取 20 个最近邻，迭代次数由 1 每次递增 1，一直到 m=10。随着迭代次数 m 的不断增加，观察算法 MAE 的变化。实验结果如图一所示，当填补系数为 0.2 时，随着迭代次数的增大，MAE 呈递减趋势，m>6 之后，MAE 的变化较小，所以迭代次数 m 为 6 时，推荐效果较好。

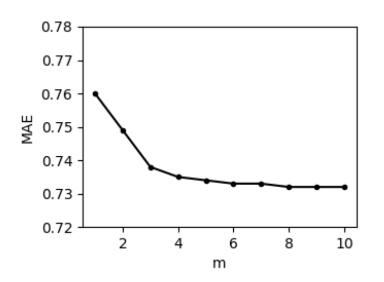


图 1 迭代次数对推荐准确度的影响

下一步确定填补系数 α 的大小。使迭代次数 m=6，改变填补系数的大小，$\alpha\in[0,1]$，每次递增 0.2。效果如图二所示，当迭代次数为 6 时，填补系数为 0.4，推荐算法效果最好。

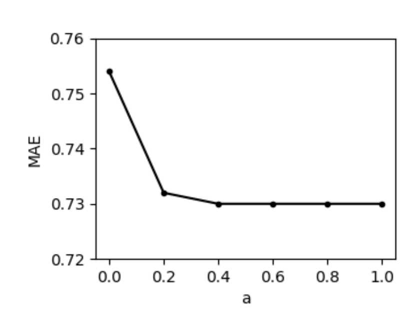


图 2 填补系数对推荐准确度的影响

第二组实验对本文算法(proposed CF)、文献[7]提出的基于项目评分预测的协同过滤推荐算法（pre-item CF）和文献[17]提出的基于改进相似度的协同过滤算法（impro-similar CF）作对

比。pre-item CF 选择余弦相似性作为用户间相似度度量方法，能达到更佳的效果[7]；本文算法迭代次数设为 6，填补系数设为 0.4。邻居个数由 5 增加到 40，实验结果如图三所示，随着邻居用户的增加，三个算法的 MAE 呈减小趋势，pre-item CF 在邻居数目大于 10 之后推荐效果稳定，但是推荐准确度不高，而 impro-similar CF 在邻居用户大于 25 之后，推荐效果稳定，本文提出的算法结合了两者的优点，推荐准确度变高，较其他两种算法表现出更好的推荐效果。

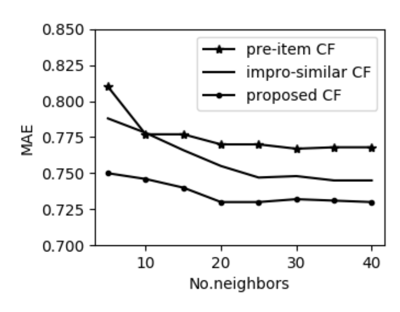


图 3　三种算法 MAE 对比图

4　结束语

针对用户评分矩阵的稀疏性问题，单一的解决办法并不能很好地解决这个问题，缺失值填补法本身是对评分缺失值的一种预测，并不能真正代表用户偏好；改进相似度方法还是在已有的有限的评分数据集上进行计算，不能从根本上解决评分数据稀疏的问题。于是，本文提出一种填补法和改进相似度相结合的结果融合的协同过滤算法。该算法首先利用填补法随机填补部分数据，填补数据量的大小由填补系数决定，然后利用改进的相似度方法进行推荐，在原始数据上再填补部分数据，进行推荐，几轮之后，统计被评分项目的平均得分，按照得分多少依次推荐给用户。本文提出的填补法和新相似性相结合的方法有更好的推荐效果，下一步工作将对本文算法两个阶段的算法，即填补法和新相似度方法作出改进。

参考文献：

[1] George G, Haas M R, Pentland A. Big data and management [J]. 吉林社会科学研究, 2014, 30 (2): 321-326.

[2] 陈洁敏，汤庸，李建国，等. 个性化推荐算法研究 [J]. 华南师范大学学报：自然科学版, 2014 (5): 8-15.

[3] Goldberg D, Oki B M, Oki B M, *et al.* Using collaborative filtering to weave an information tapestry [J]. Communications of the ACM, 1992, 35 (12): 61-70.

[4] 孙光福，吴乐，刘淇，等. 基于时序行为的协同过滤推荐算法 [J]. 软件学报, 2013, 24 (11): 2721-2733.

[5] 刘青文. 基于协同过滤的推荐算法研究 [D]. 合肥：中国科学技术大学, 2013.

[6] Deng A L, Zhu Y Y, Shi B L. A Collaborative Filtering Recommendation Algorithm Based on Item Rating Prediction [J]. Journal of Software, 2003, 14 (9): 54-65.

[7] 邓爱林，朱扬勇，施伯乐. 基于项目评分预测的协同过滤推荐算法 [J]. 软件学报, 2003, 14 (9): 1621-1628.

[8] 张玉芳，代金龙，熊忠阳. 分步填充缓解数据稀疏性的协同过滤算法 [J]. 计算机应用研究, 2013, 30 (9): 2602－2605.

[9] 吕成戍，王维国，丁永健. 基于 KNN-SVM 的混合协同过滤推荐算法 [J]. 计算机应用研究, 2012, 29 (5): 1707-1709.

[10] 冷亚军，梁昌勇，陆青，等. 基于近邻评分填补的协同过滤推荐算法 [J]. 计算机工程, 2012, 38 (21): 56-58.

[11] Chujai P, Rasmequan S, Suksawatchon U, *et al.* Imputing missing values in Collaborative Filtering using pattern frequent itemsets [C]// Proc of Electrical Engineering Congress. 2014: 1-4.

[12] Insuwan W, Suksawatchon U, Suksawatchon J. Improving missing values imputation in collaborative filtering with user-preference genre and singular value decomposition [C]// Proc of International Conference on Knowledge and Smart Technology. 2014: 87-92.

[13] 赵琴琴，鲁凯，王斌. SPCF：一种基于内存的传播式协同过滤推荐算法 [J]. 计算机学报, 2013, 36 (3): 671-676.

[14] 付芬，豆育升，韩鹏，等. 基于隐式评分和相似度传递的学习资源推荐 [J]. 计算机应用研究, 2017, 34 (12): 3725-3729.

[15] 仇国庆，马俊，赵婉滢，等. 融合正态分布函数相似度的协同过滤算法 [J/OL]. 2018, 35 (10) . [2017-09-27]. http://www. arocmag. com/article/02-2018-10-046. html.

[16] Liu H, Hu Z, Mian A, *et al.* A new user similarity model to improve the accuracy of collaborative filtering [J]. Knowledge-Based Systems, 2014, 56 (3): 156-166.

[17] 李容，李明奇，郭文强. 基于改进相似度的协同过滤算法研究 [J]. 计算机科学, 2016, 43 (12): 206-208.

[18] Sun Y, Wyk B J V, Wang Z. A new golden ratio local search based particle swarm optimization [C]// Proc of International Conference on Systems and Informatics. 2012: 754-757.

[19] Grouplens Research. MovieLens data sets [EB/OL]. (2011-08-24) . http://www. grouplens. org/node/73.